EDA

AMES HOUSING DATASET

Data Exploration, Data Cleaning , Visualization

Bhanu Pratap Singh

AmesHousingDataset

June 16, 2025

```
[ ]: # We're going to explore the "Ames Housing" dataset, which contains␣
     ↪information about houses in Ames, Iowa.
     # The dataset is available at https://www.kaggle.com/datasets/
     ↪shashanknecrothapa/ames-housing-dataset
     # We'll be learning about data exploration, cleaning, and␣
     ↪visualization using pandas and seaborn.
```

```
[ ]: # Import necessary libraries
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
```

```
[ ]: # Import the dataset
     df = pd.read_csv('C:
     ↪\\Users\\nikrc\\OneDrive\\Desktop\\Datasets\\AmesHousing.csv')
     pd.set_option('display.max_columns', None)  # To Show all columns in␣
     ↪the DataFrame
     df.head(10)
```

```
[ ]:    Order        PID  MS SubClass MS Zoning  Lot Frontage  Lot Area␣
     ↪Street  \
     0      1  526301100           20        RL         141.0     31770   Pave
     1      2  526350040           20        RH          80.0     11622   Pave
     2      3  526351010           20        RL          81.0     14267   Pave
     3      4  526353030           20        RL          93.0     11160   Pave
     4      5  527105010           60        RL          74.0     13830   Pave
     5      6  527105030           60        RL          78.0      9978   Pave
     6      7  527127150          120        RL          41.0      4920   Pave
     7      8  527145080          120        RL          43.0      5005   Pave
     8      9  527146030          120        RL          39.0      5389   Pave
     9     10  527162130           60        RL          60.0      7500   Pave

       Alley Lot Shape Land Contour Utilities Lot Config Land Slope␣
     ↪Neighborhood  \
     0   NaN       IR1          Lvl    AllPub     Corner        Gtl       ␣
     ↪NAmes
```

```
1   NaN        Reg            Lvl     AllPub     Inside         Gtl         ␣
↪NAmes
2   NaN        IR1            Lvl     AllPub     Corner         Gtl         ␣
↪NAmes
3   NaN        Reg            Lvl     AllPub     Corner         Gtl         ␣
↪NAmes
4   NaN        IR1            Lvl     AllPub     Inside         Gtl         ␣
↪Gilbert
5   NaN        IR1            Lvl     AllPub     Inside         Gtl         ␣
↪Gilbert
6   NaN        Reg            Lvl     AllPub     Inside         Gtl         ␣
↪StoneBr
7   NaN        IR1            HLS     AllPub     Inside         Gtl         ␣
↪StoneBr
8   NaN        IR1            Lvl     AllPub     Inside         Gtl         ␣
↪StoneBr
9   NaN        Reg            Lvl     AllPub     Inside         Gtl         ␣
↪Gilbert

  Condition 1 Condition 2 Bldg Type House Style  Overall Qual ␣
↪Overall Cond  \
0        Norm        Norm      1Fam      1Story            6            5
1       Feedr         Norm       1Fam       1Story            5          ␣
↪    6
2        Norm        Norm      1Fam      1Story            6            6
3        Norm        Norm      1Fam      1Story            7            5
4        Norm        Norm      1Fam      2Story            5            5
5        Norm        Norm      1Fam      2Story            6            6
6         Norm        Norm     TwnhsE       1Story            8          ␣
↪    5
7         Norm        Norm     TwnhsE       1Story            8          ␣
↪    5
8         Norm        Norm     TwnhsE       1Story            8          ␣
↪    5
9        Norm        Norm      1Fam      2Story            7            5

   Year Built  Year Remod/Add Roof Style Roof Matl Exterior 1st␣
↪Exterior 2nd  \
0         1960            1960        Hip    CompShg       BrkFace       ␣
↪Plywood
1         1961            1961      Gable    CompShg       VinylSd       ␣
↪VinylSd
2         1958            1958        Hip    CompShg       Wd Sdng       ␣
↪Wd Sdng
3         1968            1968        Hip    CompShg       BrkFace       ␣
↪BrkFace
```

```
4        1997            1998      Gable    CompShg       VinylSd       ␣
 ↪VinylSd
5        1998            1998      Gable    CompShg       VinylSd       ␣
 ↪VinylSd
6        2001            2001      Gable    CompShg       CemntBd       ␣
 ↪CmentBd
7        1992            1992      Gable    CompShg       HdBoard       ␣
 ↪HdBoard
8        1995            1996      Gable    CompShg       CemntBd       ␣
 ↪CmentBd
9        1999            1999      Gable    CompShg       VinylSd       ␣
 ↪VinylSd

  Mas Vnr Type  Mas Vnr Area Exter Qual Exter Cond Foundation Bsmt␣
 ↪Qual  \
0        Stone         112.0         TA         TA     CBlock        TA
1          NaN           0.0         TA         TA     CBlock        TA
2      BrkFace         108.0         TA         TA     CBlock        TA
3          NaN           0.0         Gd         TA     CBlock        TA
4          NaN           0.0         TA         TA      PConc        Gd
5      BrkFace          20.0         TA         TA      PConc        TA
6          NaN           0.0         Gd         TA      PConc        Gd
7          NaN           0.0         Gd         TA      PConc        Gd
8          NaN           0.0         Gd         TA      PConc        Gd
9          NaN           0.0         TA         TA      PConc        TA

  Bsmt Cond Bsmt Exposure BsmtFin Type 1  BsmtFin SF 1 BsmtFin Type␣
 ↪2  \
0        Gd            Gd            BLQ         639.0           Unf
1        TA            No            Rec         468.0           LwQ
2        TA            No            ALQ         923.0           Unf
3        TA            No            ALQ        1065.0           Unf
4        TA            No            GLQ         791.0           Unf
5        TA            No            GLQ         602.0           Unf
6        TA            Mn            GLQ         616.0           Unf
7        TA            No            ALQ         263.0           Unf
8        TA            No            GLQ        1180.0           Unf
9        TA            No            Unf           0.0           Unf

   BsmtFin SF 2  Bsmt Unf SF  Total Bsmt SF Heating Heating QC␣
 ↪Central Air  \
0           0.0        441.0         1080.0    GasA         Fa           Y
1         144.0        270.0          882.0    GasA         TA           Y
2           0.0        406.0         1329.0    GasA         TA           Y
3           0.0       1045.0         2110.0    GasA         Ex           Y
4           0.0        137.0          928.0    GasA         Gd           Y
```

```
5          0.0        324.0         926.0    GasA         Ex           Y
6          0.0        722.0        1338.0    GasA         Ex           Y
7          0.0       1017.0        1280.0    GasA         Ex           Y
8          0.0        415.0        1595.0    GasA         Ex           Y
9          0.0        994.0         994.0    GasA         Gd           Y

  Electrical  1st Flr SF  2nd Flr SF  Low Qual Fin SF  Gr Liv Area  \
0      SBrkr        1656           0                0         1656
1      SBrkr         896           0                0          896
2      SBrkr        1329           0                0         1329
3      SBrkr        2110           0                0         2110
4      SBrkr         928         701                0         1629
5      SBrkr         926         678                0         1604
6      SBrkr        1338           0                0         1338
7      SBrkr        1280           0                0         1280
8      SBrkr        1616           0                0         1616
9      SBrkr        1028         776                0         1804

   Bsmt Full Bath  Bsmt Half Bath  Full Bath  Half Bath  Bedroom␣
↪AbvGr  \
0             1.0             0.0          1          0              3
1             0.0             0.0          1          0              2
2             0.0             0.0          1          1              3
3             1.0             0.0          2          1              3
4             0.0             0.0          2          1              3
5             0.0             0.0          2          1              3
6             1.0             0.0          2          0              2
7             0.0             0.0          2          0              2
8             1.0             0.0          2          0              2
9             0.0             0.0          2          1              3

   Kitchen AbvGr Kitchen Qual  TotRms AbvGrd Functional  Fireplaces  \
0              1           TA              7        Typ           2
1              1           TA              5        Typ           0
2              1           Gd              6        Typ           0
3              1           Ex              8        Typ           2
4              1           TA              6        Typ           1
5              1           Gd              7        Typ           1
6              1           Gd              6        Typ           0
7              1           Gd              5        Typ           0
8              1           Gd              5        Typ           1
9              1           Gd              7        Typ           1

  Fireplace Qu Garage Type  Garage Yr Blt Garage Finish  Garage Cars␣
↪ \
0           Gd      Attchd         1960.0           Fin          2.0
1          NaN      Attchd         1961.0           Unf          1.0
```

```
2          NaN      Attchd         1958.0           Unf          1.0
3           TA      Attchd         1968.0           Fin          2.0
4           TA      Attchd         1997.0           Fin          2.0
5           Gd      Attchd         1998.0           Fin          2.0
6          NaN      Attchd         2001.0           Fin          2.0
7          NaN      Attchd         1992.0           RFn          2.0
8           TA      Attchd         1995.0           RFn          2.0
9           TA      Attchd         1999.0           Fin          2.0

   Garage Area Garage Qual Garage Cond Paved Drive  Wood Deck SF  \
0        528.0          TA          TA           P           210
1        730.0          TA          TA           Y           140
2        312.0          TA          TA           Y           393
3        522.0          TA          TA           Y             0
4        482.0          TA          TA           Y           212
5        470.0          TA          TA           Y           360
6        582.0          TA          TA           Y             0
7        506.0          TA          TA           Y             0
8        608.0          TA          TA           Y           237
9        442.0          TA          TA           Y           140

   Open Porch SF  Enclosed Porch  3Ssn Porch  Screen Porch  Pool␣
↪Area Pool QC  \
0             62               0           0             0          0 ␣
↪  NaN
1              0               0           0           120          0 ␣
↪  NaN
2             36               0           0             0          0 ␣
↪  NaN
3              0               0           0             0          0 ␣
↪  NaN
4             34               0           0             0          0 ␣
↪  NaN
5             36               0           0             0          0 ␣
↪  NaN
6              0             170           0             0          0 ␣
↪  NaN
7             82               0           0           144          0 ␣
↪  NaN
8            152               0           0             0          0 ␣
↪  NaN
9             60               0           0             0          0 ␣
↪  NaN

  Fence Misc Feature  Misc Val  Mo Sold  Yr Sold Sale Type Sale␣
↪Condition  \
```

```
0    NaN          NaN       0       5     2010       WD         Normal
1  MnPrv          NaN       0       6     2010       WD         Normal
2    NaN         Gar2   12500       6     2010       WD         Normal
3    NaN          NaN       0       4     2010       WD         Normal
4  MnPrv          NaN       0       3     2010       WD         Normal
5    NaN          NaN       0       6     2010       WD         Normal
6    NaN          NaN       0       4     2010       WD         Normal
7    NaN          NaN       0       1     2010       WD         Normal
8    NaN          NaN       0       3     2010       WD         Normal
9    NaN          NaN       0       6     2010       WD         Normal

   SalePrice
0     215000
1     105000
2     172000
3     244000
4     189900
5     195500
6     213500
7     191500
8     236500
9     189000
```

[ ]:
```
# Why do we use this?
# To suppress FutureWarnings that may arise from using deprecated␣
↪features in libraries like pandas or numpy.
# This is useful to keep the output clean, especially when running␣
↪scripts that may generate many warnings.

import warnings
warnings.filterwarnings('ignore',category=FutureWarning)
```

[12]:
```
# Check the shape of the Dataset
print(f"The dataset has {df.shape[0]} rows and {df.shape[1]}␣
↪columns.")
```

```
The dataset has 2930 rows and 82 columns.
```

[13]:
```
# View Columns types and non-null counts
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2930 entries, 0 to 2929
Data columns (total 82 columns):
 #    Column           Non-Null Count   Dtype
---   ------           --------------   -----
 0    Order            2930 non-null    int64
 1    PID              2930 non-null    int64
```

```
2   MS SubClass      2930 non-null   int64
3   MS Zoning        2930 non-null   object
4   Lot Frontage     2440 non-null   float64
5   Lot Area         2930 non-null   int64
6   Street           2930 non-null   object
7   Alley            198 non-null    object
8   Lot Shape        2930 non-null   object
9   Land Contour     2930 non-null   object
10  Utilities        2930 non-null   object
11  Lot Config       2930 non-null   object
12  Land Slope       2930 non-null   object
13  Neighborhood     2930 non-null   object
14  Condition 1      2930 non-null   object
15  Condition 2      2930 non-null   object
16  Bldg Type        2930 non-null   object
17  House Style      2930 non-null   object
18  Overall Qual     2930 non-null   int64
19  Overall Cond     2930 non-null   int64
20  Year Built       2930 non-null   int64
21  Year Remod/Add   2930 non-null   int64
22  Roof Style       2930 non-null   object
23  Roof Matl        2930 non-null   object
24  Exterior 1st     2930 non-null   object
25  Exterior 2nd     2930 non-null   object
26  Mas Vnr Type     1155 non-null   object
27  Mas Vnr Area     2907 non-null   float64
28  Exter Qual       2930 non-null   object
29  Exter Cond       2930 non-null   object
30  Foundation       2930 non-null   object
31  Bsmt Qual        2850 non-null   object
32  Bsmt Cond        2850 non-null   object
33  Bsmt Exposure    2847 non-null   object
34  BsmtFin Type 1   2850 non-null   object
35  BsmtFin SF 1     2929 non-null   float64
36  BsmtFin Type 2   2849 non-null   object
37  BsmtFin SF 2     2929 non-null   float64
38  Bsmt Unf SF      2929 non-null   float64
39  Total Bsmt SF    2929 non-null   float64
40  Heating          2930 non-null   object
41  Heating QC       2930 non-null   object
42  Central Air      2930 non-null   object
43  Electrical       2929 non-null   object
44  1st Flr SF       2930 non-null   int64
45  2nd Flr SF       2930 non-null   int64
46  Low Qual Fin SF  2930 non-null   int64
47  Gr Liv Area      2930 non-null   int64
48  Bsmt Full Bath   2928 non-null   float64
49  Bsmt Half Bath   2928 non-null   float64
```

```
 50  Full Bath        2930 non-null   int64
 51  Half Bath        2930 non-null   int64
 52  Bedroom AbvGr    2930 non-null   int64
 53  Kitchen AbvGr    2930 non-null   int64
 54  Kitchen Qual     2930 non-null   object
 55  TotRms AbvGrd    2930 non-null   int64
 56  Functional       2930 non-null   object
 57  Fireplaces       2930 non-null   int64
 58  Fireplace Qu     1508 non-null   object
 59  Garage Type      2773 non-null   object
 60  Garage Yr Blt    2771 non-null   float64
 61  Garage Finish    2771 non-null   object
 62  Garage Cars      2929 non-null   float64
 63  Garage Area      2929 non-null   float64
 64  Garage Qual      2771 non-null   object
 65  Garage Cond      2771 non-null   object
 66  Paved Drive      2930 non-null   object
 67  Wood Deck SF     2930 non-null   int64
 68  Open Porch SF    2930 non-null   int64
 69  Enclosed Porch   2930 non-null   int64
 70  3Ssn Porch       2930 non-null   int64
 71  Screen Porch     2930 non-null   int64
 72  Pool Area        2930 non-null   int64
 73  Pool QC          13 non-null     object
 74  Fence            572 non-null    object
 75  Misc Feature     106 non-null    object
 76  Misc Val         2930 non-null   int64
 77  Mo Sold          2930 non-null   int64
 78  Yr Sold          2930 non-null   int64
 79  Sale Type        2930 non-null   object
 80  Sale Condition   2930 non-null   object
 81  SalePrice        2930 non-null   int64
dtypes: float64(11), int64(28), object(43)
memory usage: 1.8+ MB
None
```

[ ]:
```
# Descriptive statistics of the dataset
# This provides a summary of the central tendency, dispersion, and
  ↪shape of the dataset's distribution,
df.describe().T
```

[ ]:
```
                 count          mean           std          min  \
Order           2930.0  1.465500e+03  8.459625e+02          1.0
PID             2930.0  7.144645e+08  1.887308e+08  526301100.0
MS SubClass     2930.0  5.738737e+01  4.263802e+01         20.0
Lot Frontage    2440.0  6.922459e+01  2.336533e+01         21.0
Lot Area        2930.0  1.014792e+04  7.880018e+03       1300.0
```

```
Overall Qual     2930.0  6.094881e+00  1.411026e+00        1.0
Overall Cond     2930.0  5.563140e+00  1.111537e+00        1.0
Year Built       2930.0  1.971356e+03  3.024536e+01     1872.0
Year Remod/Add   2930.0  1.984267e+03  2.086029e+01     1950.0
Mas Vnr Area     2907.0  1.018968e+02  1.791126e+02        0.0
BsmtFin SF 1     2929.0  4.426296e+02  4.555908e+02        0.0
BsmtFin SF 2     2929.0  4.972243e+01  1.691685e+02        0.0
Bsmt Unf SF      2929.0  5.592625e+02  4.394942e+02        0.0
Total Bsmt SF    2929.0  1.051615e+03  4.406151e+02        0.0
1st Flr SF       2930.0  1.159558e+03  3.918909e+02      334.0
2nd Flr SF       2930.0  3.354560e+02  4.283957e+02        0.0
Low Qual Fin SF  2930.0  4.676792e+00  4.631051e+01        0.0
Gr Liv Area      2930.0  1.499690e+03  5.055089e+02      334.0
Bsmt Full Bath   2928.0  4.313525e-01  5.248202e-01        0.0
Bsmt Half Bath   2928.0  6.113388e-02  2.452536e-01        0.0
Full Bath        2930.0  1.566553e+00  5.529406e-01        0.0
Half Bath        2930.0  3.795222e-01  5.026293e-01        0.0
Bedroom AbvGr    2930.0  2.854266e+00  8.277311e-01        0.0
Kitchen AbvGr    2930.0  1.044369e+00  2.140762e-01        0.0
TotRms AbvGrd    2930.0  6.443003e+00  1.572964e+00        2.0
Fireplaces       2930.0  5.993174e-01  6.479209e-01        0.0
Garage Yr Blt    2771.0  1.978132e+03  2.552841e+01     1895.0
Garage Cars      2929.0  1.766815e+00  7.605664e-01        0.0
Garage Area      2929.0  4.728197e+02  2.150465e+02        0.0
Wood Deck SF     2930.0  9.375188e+01  1.263616e+02        0.0
Open Porch SF    2930.0  4.753345e+01  6.748340e+01        0.0
Enclosed Porch   2930.0  2.301160e+01  6.413906e+01        0.0
3Ssn Porch       2930.0  2.592491e+00  2.514133e+01        0.0
Screen Porch     2930.0  1.600205e+01  5.608737e+01        0.0
Pool Area        2930.0  2.243345e+00  3.559718e+01        0.0
Misc Val         2930.0  5.063515e+01  5.663443e+02        0.0
Mo Sold          2930.0  6.216041e+00  2.714492e+00        1.0
Yr Sold          2930.0  2.007790e+03  1.316613e+00     2006.0
SalePrice        2930.0  1.807961e+05  7.988669e+04    12789.0

                          25%          50%           75%           max
Order            7.332500e+02       1465.5  2.197750e+03  2.930000e+03
PID              5.284770e+08  535453620.0  9.071811e+08  1.007100e+09
MS SubClass      2.000000e+01         50.0  7.000000e+01  1.900000e+02
Lot Frontage     5.800000e+01         68.0  8.000000e+01  3.130000e+02
Lot Area         7.440250e+03       9436.5  1.155525e+04  2.152450e+05
Overall Qual     5.000000e+00          6.0  7.000000e+00  1.000000e+01
Overall Cond     5.000000e+00          5.0  6.000000e+00  9.000000e+00
Year Built       1.954000e+03       1973.0  2.001000e+03  2.010000e+03
Year Remod/Add   1.965000e+03       1993.0  2.004000e+03  2.010000e+03
Mas Vnr Area     0.000000e+00          0.0  1.640000e+02  1.600000e+03
BsmtFin SF 1     0.000000e+00        370.0  7.340000e+02  5.644000e+03
```

```
BsmtFin SF 2     0.000000e+00         0.0  0.000000e+00  1.526000e+03
Bsmt Unf SF      2.190000e+02       466.0  8.020000e+02  2.336000e+03
Total Bsmt SF    7.930000e+02       990.0  1.302000e+03  6.110000e+03
1st Flr SF       8.762500e+02      1084.0  1.384000e+03  5.095000e+03
2nd Flr SF       0.000000e+00         0.0  7.037500e+02  2.065000e+03
Low Qual Fin SF  0.000000e+00         0.0  0.000000e+00  1.064000e+03
Gr Liv Area      1.126000e+03      1442.0  1.742750e+03  5.642000e+03
Bsmt Full Bath   0.000000e+00         0.0  1.000000e+00  3.000000e+00
Bsmt Half Bath   0.000000e+00         0.0  0.000000e+00  2.000000e+00
Full Bath        1.000000e+00         2.0  2.000000e+00  4.000000e+00
Half Bath        0.000000e+00         0.0  1.000000e+00  2.000000e+00
Bedroom AbvGr    2.000000e+00         3.0  3.000000e+00  8.000000e+00
Kitchen AbvGr    1.000000e+00         1.0  1.000000e+00  3.000000e+00
TotRms AbvGrd    5.000000e+00         6.0  7.000000e+00  1.500000e+01
Fireplaces       0.000000e+00         1.0  1.000000e+00  4.000000e+00
Garage Yr Blt    1.960000e+03      1979.0  2.002000e+03  2.207000e+03
Garage Cars      1.000000e+00         2.0  2.000000e+00  5.000000e+00
Garage Area      3.200000e+02       480.0  5.760000e+02  1.488000e+03
Wood Deck SF     0.000000e+00         0.0  1.680000e+02  1.424000e+03
Open Porch SF    0.000000e+00        27.0  7.000000e+01  7.420000e+02
Enclosed Porch   0.000000e+00         0.0  0.000000e+00  1.012000e+03
3Ssn Porch       0.000000e+00         0.0  0.000000e+00  5.080000e+02
Screen Porch     0.000000e+00         0.0  0.000000e+00  5.760000e+02
Pool Area        0.000000e+00         0.0  0.000000e+00  8.000000e+02
Misc Val         0.000000e+00         0.0  0.000000e+00  1.700000e+04
Mo Sold          4.000000e+00         6.0  8.000000e+00  1.200000e+01
Yr Sold          2.007000e+03      2008.0  2.009000e+03  2.010000e+03
SalePrice        1.295000e+05    160000.0  2.135000e+05  7.550000e+05
```

[28]:
```python
# Total Missing Values per Column
df.isnull().sum().sort_values(ascending=False).head(20)
```

[28]:
```
Pool QC           2917
Misc Feature      2824
Alley             2732
Fence             2358
Mas Vnr Type      1775
Fireplace Qu      1422
Lot Frontage       490
Garage Qual        159
Garage Yr Blt      159
Garage Cond        159
Garage Finish      159
Garage Type        157
Bsmt Exposure       83
BsmtFin Type 2      81
Bsmt Qual           80
```

```
Bsmt Cond          80
BsmtFin Type 1     80
Mas Vnr Area       23
Bsmt Full Bath      2
Bsmt Half Bath      2
dtype: int64
```

[32]:
```
df['Lot Frontage'].fillna(df['Lot Frontage'].mean(),inplace=True)
```

[33]:
```
df['Alley'].fillna(df['Alley'].mode()[0],inplace=True)
```

[54]:
```
# Total Missing Values per Column
missing_values = df.isnull().sum()/len(df) * 100
print("Percentage of Missing Values per Column:")
print(missing_values[missing_values > 0].
  ↪sort_values(ascending=False))
```

```
Percentage of Missing Values per Column:
Series([], dtype: float64)
```

[38]:
```
#  Dropping Columns with more than 50% missing values
threshold = 0.5
cols = df.columns[missing_values > threshold]
df.drop(cols, axis=1, inplace=True)
```

[ ]:
```
# Mode fill for discrete(categorical) features
mode_cols = ['Bsmt Half Bath', 'Bsmt Full Bath', 'Electrical']
for col in mode_cols:
    df[col].fillna(df[col].mode()[0], inplace=True)

# Median fill for numeric features
median_cols = ['BsmtFin SF 1', 'BsmtFin SF 2', 'Bsmt Unf SF', 'Total␣
  ↪Bsmt SF', 'Garage Cars', 'Garage Area']
for col in median_cols:
    df[col].fillna(df[col].median(), inplace=True)
```

[59]:
```
# Checking the percentage of missing values after filling
print("Percentage of Missing Values after filling:")
(df.isnull().sum() / len(df)) * 100
```

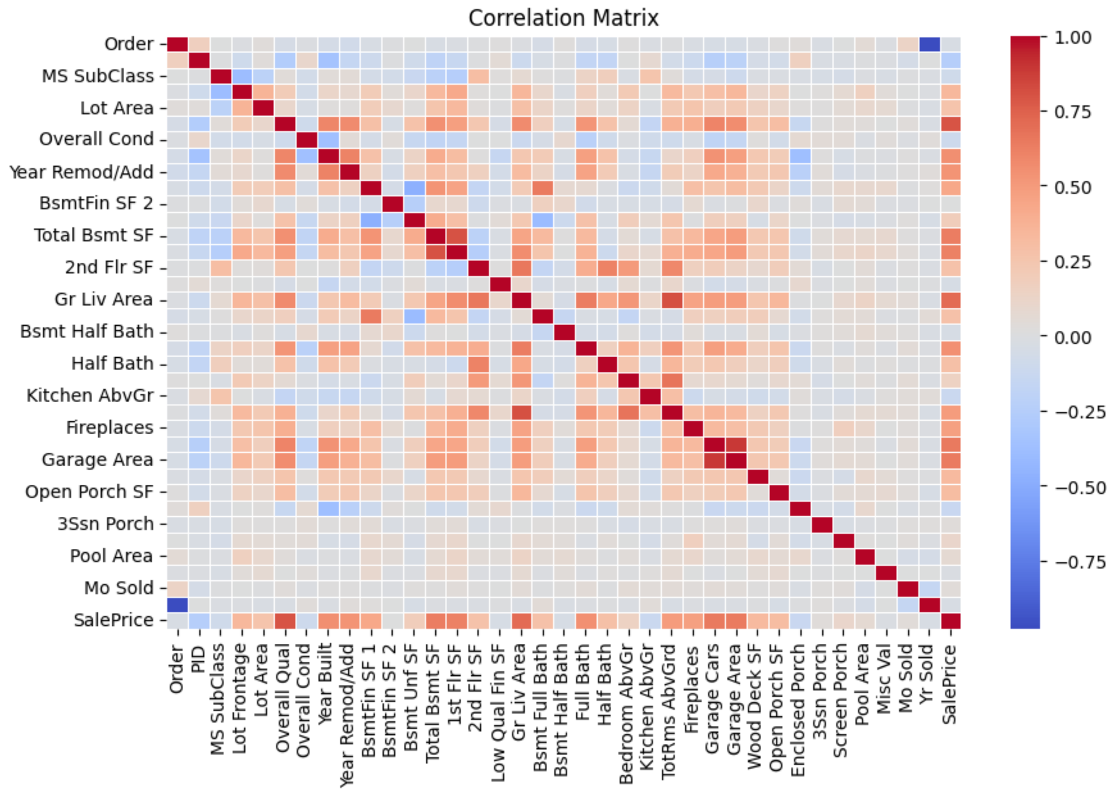
```
Percentage of Missing Values after filling:
```

[59]:
```
Order            0.0
PID              0.0
MS SubClass      0.0
MS Zoning        0.0
Lot Frontage     0.0
```

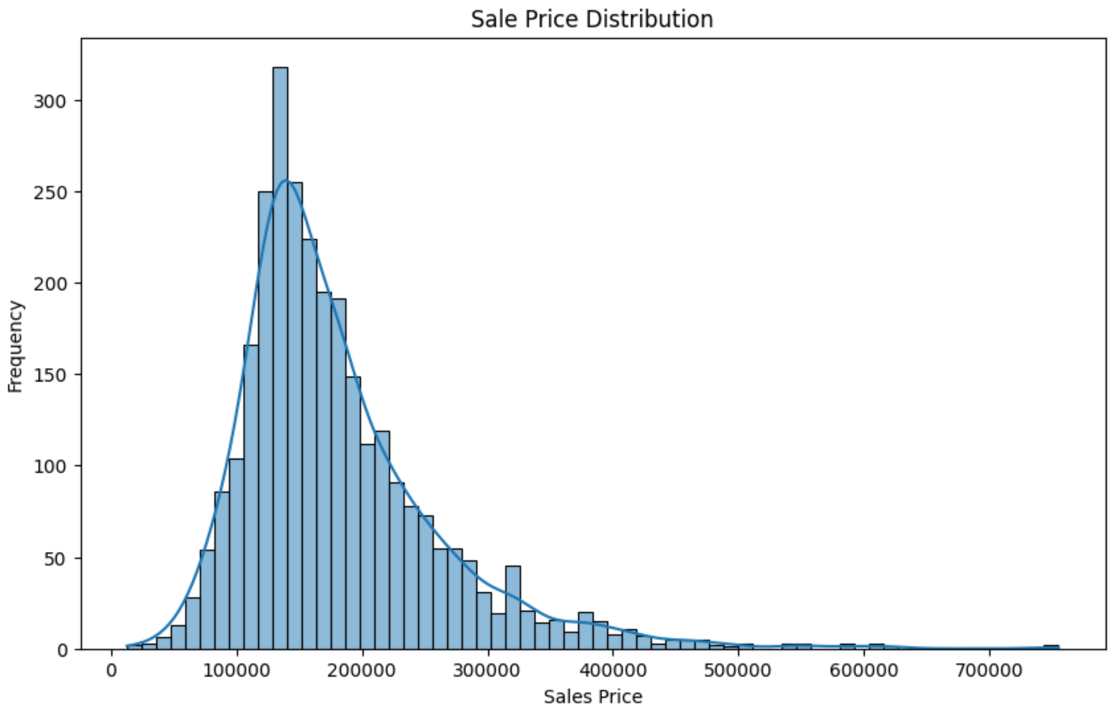
```
                      …
Mo Sold             0.0
Yr Sold             0.0
Sale Type           0.0
Sale Condition      0.0
SalePrice           0.0
Length: 66, dtype: float64
```

[ ]:
```
# What is Correlation Matrix ?
# The correlation matrix is a table that shows the correlation␣
  ↪coefficients between many variables.
# Each cell in the table displays the correlation between two␣
  ↪variables.
# The value is between -1 and 1. A value closer to 1 means a strong␣
  ↪positive correlation, while a value closer to -1 means a strong␣
  ↪negative correlation.
```

[67]:
```
# Correlation Matrix

plt.figure(figsize = (10,6))

# Selecting only numeric data for correlation matrix
numeric_data = df.select_dtypes(include = [np.number])
# Note: numeric_data.corr(): It returns a correlation matrix of only␣
  ↪the numeric columns in your DataFrame.
sns.heatmap(numeric_data.corr(),cmap =␣
  ↪"coolwarm",annot=False,linewidths=0.5)
plt.title("Correlation Matrix")

plt.show()
```



[72]:
```
# Price Distribution
plt.figure(figsize=(10, 6))
sns.histplot(df['SalePrice'],kde = True)
plt.title("Sale Price Distribution")
plt.xlabel("Sales Price")
plt.ylabel("Frequency")
plt.show()
```



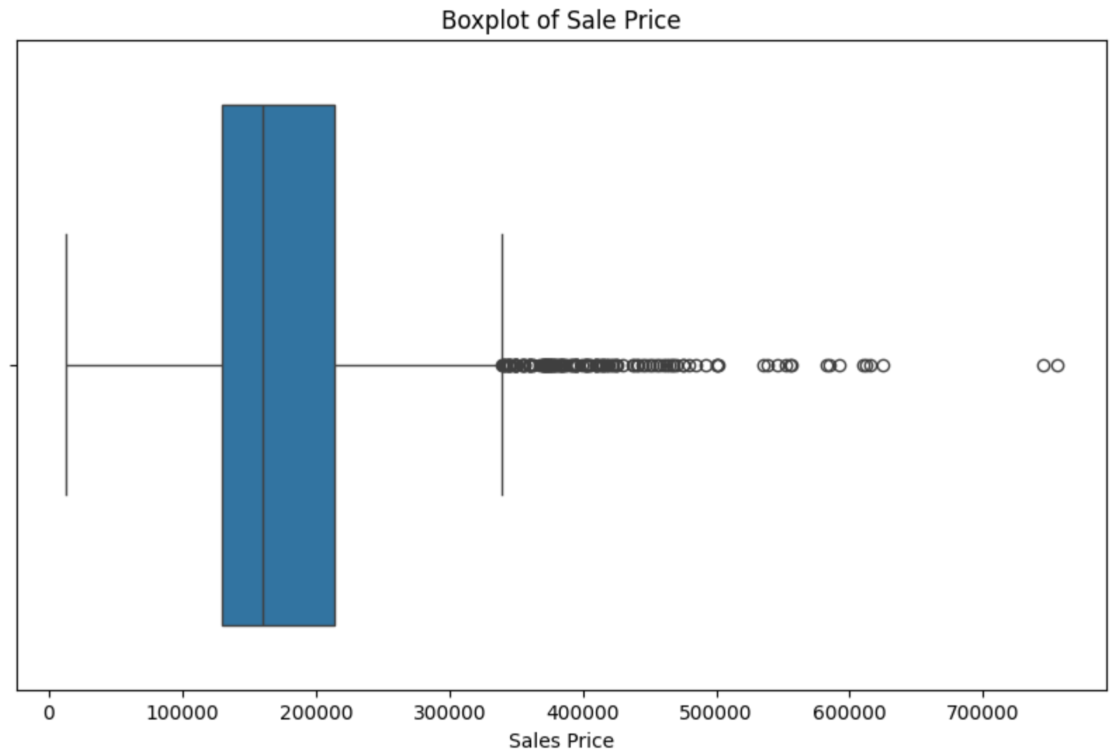
[ ]:
```
# Clearly the distribution is right skewed, indicating that most
↪houses are sold at lower prices, with fewer houses sold at higher
↪prices.
# There are a few outliers on the higher end of the price spectrum,
↪which is common in real estate data.
```

[ ]:
```
# There's a way to find outliers using the Interquartile Range (IQR)
↪method.
# Steps to find outliers using IQR:
# 1. Calculate the first quartile (Q1) and third quartile (Q3) of
↪the data.
# 2. Compute the interquartile range (IQR) as Q3 - Q1.
# 3. Determine the lower bound as Q1 - 1.5 * IQR and the upper bound
↪as Q3 + 1.5 * IQR.
# 4. Any data point outside these bounds is considered an outlier.

def find_outliers_iqr(data):
    Q1 = np.percentile(data,25)
    Q3 = np.percentile(data,75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = data[(data < lower_bound) | (data > upper_bound)]
```

```
    return outliers

# Finding outliers in SalePrice
outliers = find_outliers_iqr(df['SalePrice'])
print(f"Number of outliers in SalePrice: {len(outliers)}")


# Note:
# What to do with outliers?
# 1. Remove them: If they are errors or not relevant to the analysis.
# 2. Transform them: Use transformations like log or square root to␣
  ↪reduce their impact.
# 3. Keep them: If they are valid observations that provide␣
  ↪important information.

# For this dataset, we will keep the outliers as they may represent␣
  ↪high-value properties that are important for analysis.
```

```
Number of outliers in SalePrice: 137
```

[79]:
```
# Visualizing Outliers
# Why do we visualize outliers?
# Visualizing outliers helps to understand their distribution and␣
  ↪impact on the dataset.
# It allows us to see how they affect the overall analysis and␣
  ↪whether they should be treated differently.

plt.figure(figsize=(10, 6))
sns.boxplot(x=df['SalePrice'])
plt.title("Boxplot of Sale Price")
plt.xlabel("Sales Price")
plt.show()
```



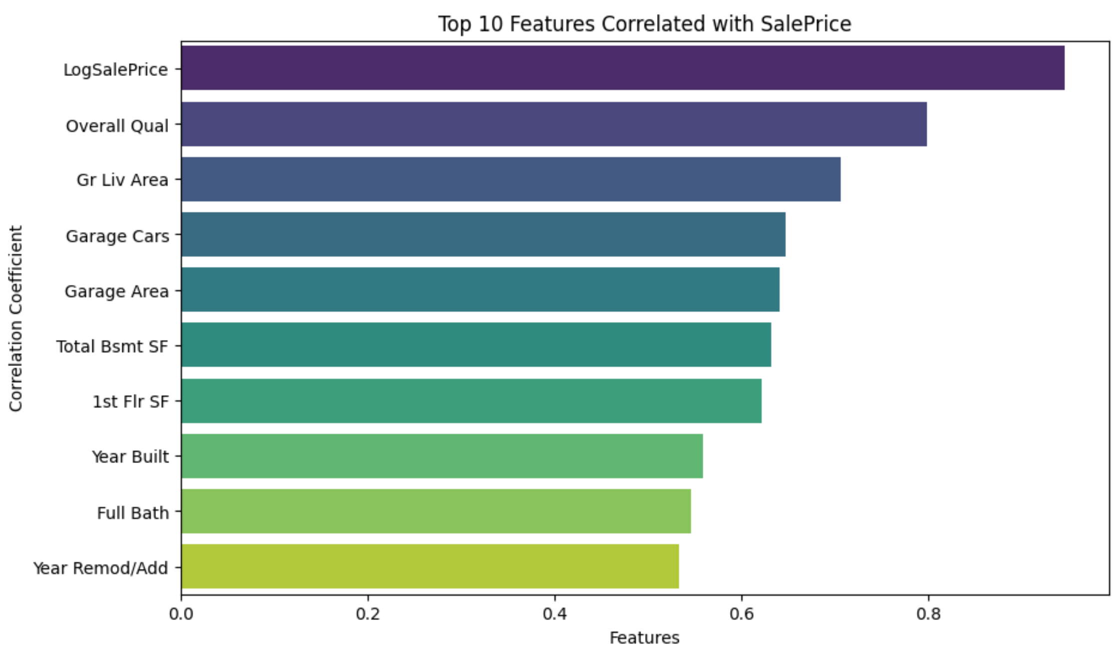
[ ]:
```
# What can we conclude from the boxplot?
# The boxplot shows that the majority of the data is concentrated in
↪the lower price range, with a few high-value outliers.
# The whiskers extend to the lower and upper bounds, while the
↪outliers are represented as individual points beyond these bounds.
# This indicates that while most houses are sold at lower prices,
↪there are a few high-value properties that significantly impact
↪the average price.
```

[90]:
```
# Finding highly correlated features with SalePrice
numeric_df = df.select_dtypes(include=[np.number])
correlation_matrix = numeric_df.corr()

# Selecting features with correlation greater than 0.5 with SalePrice
highly_correlated_features =
↪correlation_matrix['SalePrice'][abs(correlation_matrix['SalePrice'])
↪> 0.5].index.tolist()
print("Features highly correlated with SalePrice:")
print(highly_correlated_features)
```

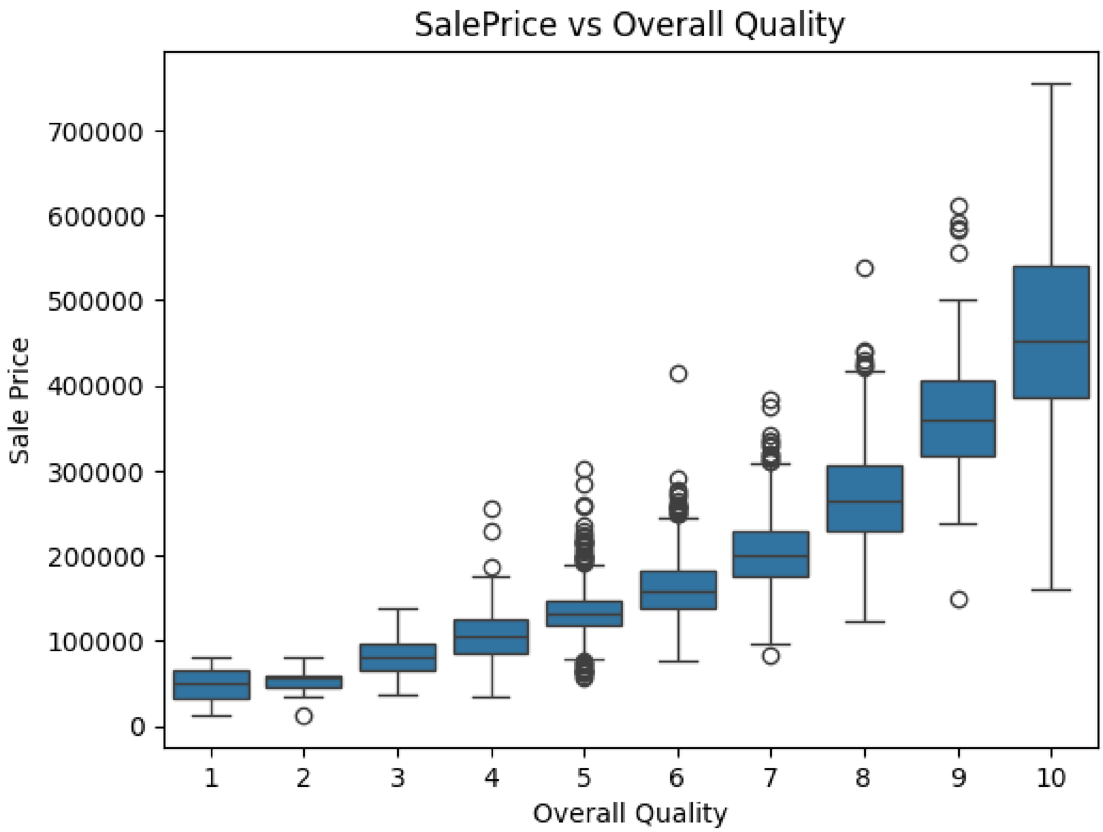
```
Features highly correlated with SalePrice:
['Overall Qual', 'Year Built', 'Year Remod/Add', 'Total Bsmt SF',
↪'1st Flr SF',
```

'Gr Liv Area', 'Full Bath', 'Garage Cars', 'Garage Area', 'SalePrice',
'LogSalePrice']

[ ]:
```
#  Visualizing the relationship between SalePrice and highly␣
↪correlated features

# Sorting and plotting the top 10 features
corr_matrix = correlation_matrix['SalePrice'].drop('SalePrice')
top_10_features = corr_matrix.abs().sort_values(ascending=False).
↪head(10)
plt.figure(figsize = (10,6))
sns.barplot(x = top_10_features.values, y = top_10_features.index,␣
↪palette='viridis')
plt.title("Top 10 Features Correlated with SalePrice")
plt.xlabel("Features")
plt.ylabel("Correlation Coefficient")
plt.show()
```



[ ]:
```
# How does Overall Qual affect SalePrice?
sns.boxplot(x='Overall Qual', y='SalePrice', data=df)
plt.title("SalePrice vs Overall Quality")
plt.xlabel("Overall Quality")
plt.ylabel("Sale Price")
plt.show()
```

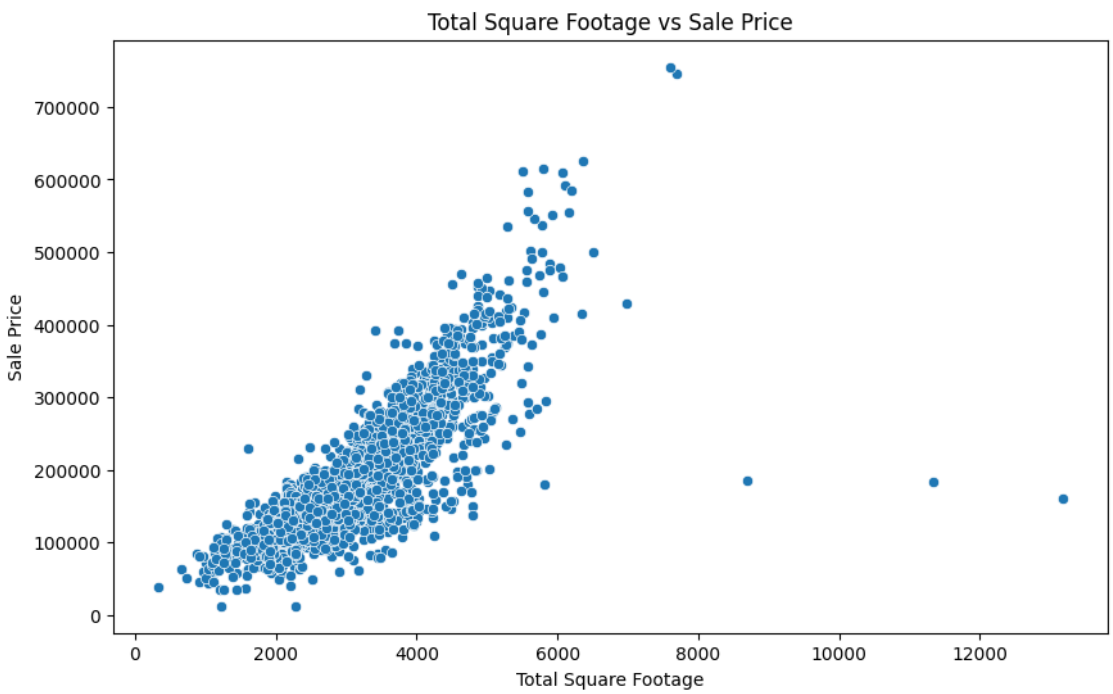



```
[ ]: # What can we conclude from this boxplot?
     # The boxplot shows that as the overall quality of the house␣
      ↪increases, the sale price also tends to increase.
     # Higher quality houses (with higher Overall Qual ratings) have a␣
      ↪wider range of sale prices, indicating that they are generally␣
      ↪more expensive.
     # This suggests that overall quality is a significant factor in␣
      ↪determining the sale price of a house.
```

```
[103]: # We can also create new features based on existing ones to enhance␣
        ↪our analysis.


       # Total bathrooms
       df['Total Bathrooms'] = df['Full Bath'] + df['Half Bath'] + df['Bsmt␣
        ↪Full Bath'] + df['Bsmt Half Bath']

       # Total Square Footage
       df['Total Square Footage'] = df['Gr Liv Area'] + df['Total Bsmt SF']␣
        ↪+ df['Garage Area']
```

```
# House Age
df['House Age'] = df['Yr Sold'] - df['Year Built']

# IsRemodelled
df['IsRemodelled'] = (df['Year Remod/Add'] != df['Year Built']).
  ↪astype(int)

# Price per Square Foot
df['Price per Sq Ft'] = df['SalePrice'] / df['Total Square Footage']
```

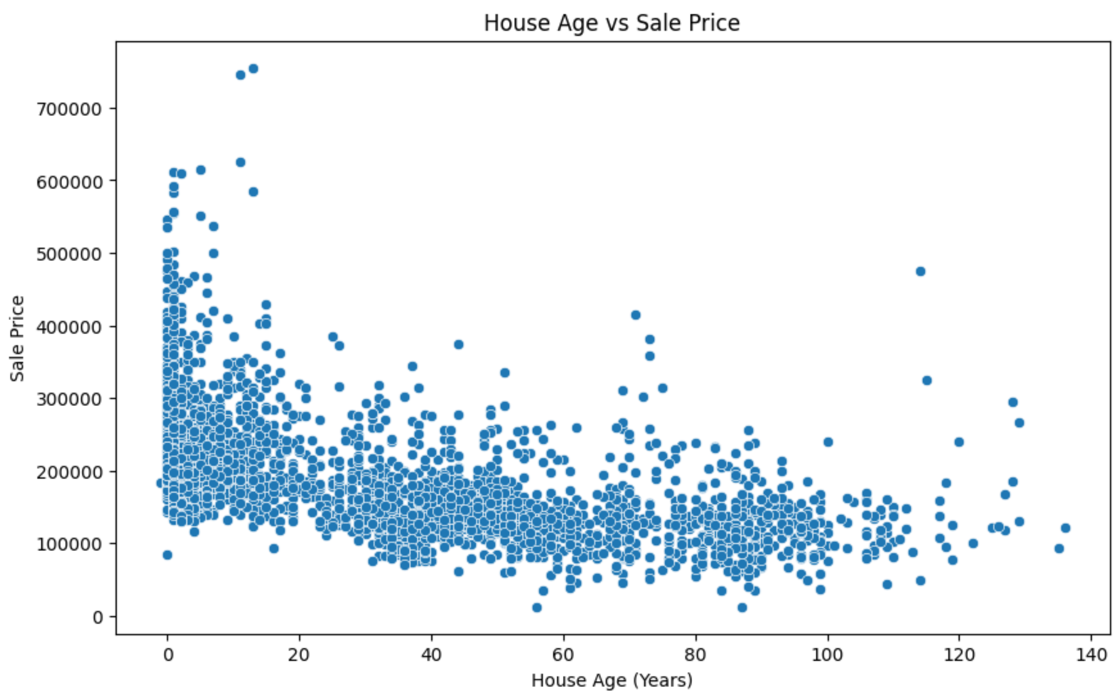
[104]:
```
# Visualizing the new features
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Total Square Footage', y='SalePrice', data=df)
plt.title("Total Square Footage vs Sale Price")
plt.xlabel("Total Square Footage")
plt.ylabel("Sale Price")
plt.show()
```



[ ]:
```
# What can we conclude from the boxplot?
# The scatter plot shows a positive correlation between total square␣
  ↪footage and sale price.
# As the total square footage of the house increases, the sale price␣
  ↪also tends to increase.
```

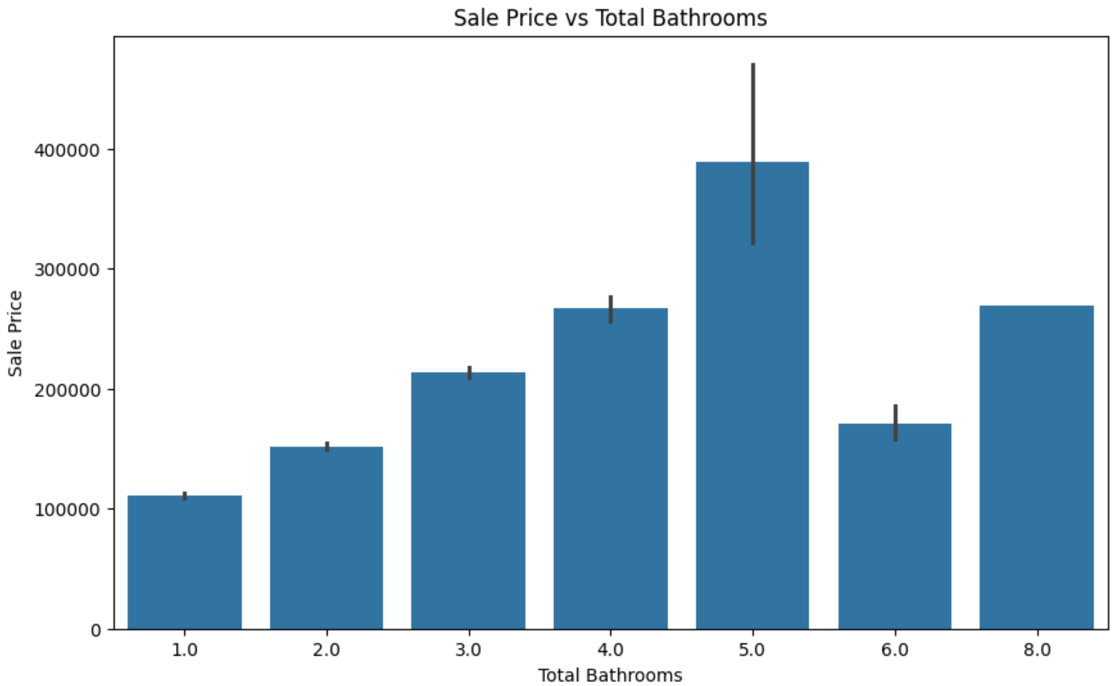
```
# This suggests that larger houses generally command higher prices␣
↪in the real estate market.And ofcourse there are exceptions.
```

[105]:
```
# Visualizing the relationship between House Age and Sale Price
plt.figure(figsize=(10, 6))
sns.scatterplot(x='House Age', y='SalePrice', data=df)
plt.title("House Age vs Sale Price")
plt.xlabel("House Age (Years)")
plt.ylabel("Sale Price")
plt.show()
```



[ ]:
```
# What can we conclude from the boxplot?
# The scatter plot shows that there is a general trend where older␣
↪houses tend to have lower sale prices.
```

[109]:
```
# Visualizing if the number of bathrooms affects the Sale Price
plt.figure(figsize=(10, 6))
sns.barplot(x='Total Bathrooms', y='SalePrice', data=df)
plt.title("Sale Price vs Total Bathrooms")
plt.xlabel("Total Bathrooms")
plt.ylabel("Sale Price")
plt.show()
```



[ ]:
```
# What can we conclude from the barplot?
# The bar plot shows that as the number of total bathrooms␣
↪increases, the sale price also tends to increase.
```

[ ]:
```
# This is where we end the exploratory data analysis (EDA) for the␣
↪Ames Housing dataset.
# We have explored the dataset, handled missing values, visualized␣
↪distributions, identified outliers, and created new features.
# Hope this analysis helps you understand the dataset better and␣
↪provides insights into the factors affecting house prices in␣
↪Ames, Iowa.
# And I do hope if you could suggest any improvements or additional␣
↪analyses that could be performed on this dataset, it would be␣
↪greatly appreciated.
```